Machine Learning 2 - Homework Assignment 3

Alexandra Lindt

October 2, 2019

Problem 1

1.

$$\begin{aligned}
H(X,Y) &= -\int p(x,y)\log(p(x,y))dxdy \\
\star &= -\int p(x,y)\log(p(x|y)p(y))dxdy \\
&= -\int p(x,y)\log(p(x|y))dxdy - \int p(x,y)\log(p(y))dxdy \\
&= H(X|Y) - \int p(y)\log(p(y))dy \int p(x|y)dx \\
&= H(X|Y) - \int p(y)\log(p(y))dy \\
&= H(X|Y) + H(Y)
\end{aligned}$$

$\star$: If we use $p(x,y) = p(y|x)p(x)$ instead of $p(x,y) = p(x|y)p(y)$ here, we do obtain $H(X,Y) = H(Y|X) + H(X)$ in analogy to the given derivation.

2.

$$\begin{aligned}
I[X,Y|Z] &= -\iiint p(z)p(x,y|z)\log(\frac{p(x|z)p(y|z)}{p(x,y|z)})dxdydz \\
\star &= -\iiint p(z)p(x,y|z)\log(\frac{p(x|z)}{p(x|y,z)})dxdydz \\
&= -\iiint p(z)p(x,y|z)\Big(\log(p(x|z)) - \log(p(x|y,z))\Big)dxdydz \\
&= -\iiint p(z)p(x,y|z)\log(p(x|z))dxdydz + \iiint p(z)p(x,y|z)\log(p(x|y,z))dxdydz \\
&= -\iiint p(x,y,z)\log(p(x|z))dxdydz + \iiint p(x,y,z)\log(p(x|y,z))dxdydz
\end{aligned}$$

$$
\begin{aligned}
&= -\iint \log(p(x|z))dxdz \int p(x,z|y)p(y)dy - H(X|Y,Z) \\
&= -\iint p(x,z)\log(p(x|z))dxdz - H(X|Y,Z) \\
&= H(X|Z) - H(X|Y,Z)
\end{aligned}
$$

$\star$: If we use $\frac{p(x|z)p(y|z)}{p(x,y|z)} = \frac{p(y|z)}{p(y|x,z)}$ instead of $\frac{p(x|z)p(y|z)}{p(x,y|z)} = \frac{p(x|z)}{p(x|y,z)}$ here, we do obtain $I[X,Y|Z] = H(Y|Z) - H(Y|X,Z)$ in analogy to the given derivation.

Problem 2

1.

$$
\begin{aligned}
\text{Mult}(x|\pi) &= \frac{M!}{\prod_{i=1}^K x_i} \prod_{i=1}^K \pi_i^{x_i} \\
&= \frac{M!}{\prod_{i=1}^K x_i} \exp\big(\log \prod_{i=1}^K \pi_i^{x_i}\big) \\
&= \frac{M!}{\prod_{i=1}^K x_i} \exp\big(\sum_{i=1}^K x_i \log(\pi_i)\big) \\
&= \frac{M!}{\prod_{i=1}^K x_i} \exp\big(\sum_{i=1}^{K-1} x_i \log(\pi_i) + x_K \log(\pi_K)\big) \\
&= \frac{M!}{\prod_{i=1}^K x_i} \exp\big(\sum_{i=1}^{K-1} x_i \log(\pi_i) + (M - \sum_{i=1}^{K-1} x_i)\log(1 - \sum_{i=1}^{K-1}\pi_i)\big) \\
&= \frac{M!}{\prod_{i=1}^K x_i} \exp\big(\sum_{i=1}^{K-1} x_i \log(\pi_i) - \sum_{i=1}^{K-1} x_i \log(1 - \sum_{i=1}^{K-1}\pi_i) + M\log(1 - \sum_{i=1}^{K-1}\pi_i)\big) \\
&= \frac{M!}{\prod_{i=1}^K x_i} \exp\big(\sum_{i=1}^{K-1} x_i \log\left(\frac{\pi_i}{1-\sum_{i=1}^{K-1}\pi_i}\right) + M\log(1 - \sum_{i=1}^{K-1}\pi_i)\big)
\end{aligned}
$$

$$
\implies T(x) = \frac{M!}{\prod_{i=1}^K x_i} \qquad \eta = \begin{bmatrix} \log\frac{\pi_1}{1-\sum_{i=1}^{K-1}\pi_i} \\ \vdots \\ \log\frac{\pi_{K-1}}{1-\sum_{i=1}^{K-1}\pi_i} \\ 0 \end{bmatrix} \qquad u(x) = \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix}
$$

$$
\begin{aligned}
\implies \mathcal{A}(\eta) &= -M\log(1 - \sum_{i=1}^{K-1}\pi_i) = M\log\left(\frac{1}{1-\sum_{i=1}^{K-1}\pi_i}\right) = M\log\left(\frac{1-1+\sum_{i=1}^{K-1}\pi_i}{1-\sum_{i=1}^{K-1}\pi_i} + 1\right) \\
&= M\log\left(1 + \sum_{i=1}^{K-1}\frac{\pi_i}{1-\sum_{i=1}^{K-1}\pi_i}\right) = M\log(1 + \sum_{i=1}^{K-1}\exp(\eta_i))
\end{aligned}
$$

2.

$$\mathbb{E}(x_i) = \frac{\partial \mathcal{A}(\eta)}{\partial \eta_i} = \frac{\partial}{\partial \eta_i} M \log(1 + \sum_{i=1}^{K-1} \exp(\eta_i)) = M \frac{\exp(\eta_i)}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)}$$

$$= M \frac{\frac{\pi_i}{1-\sum_{i=1}^{K-1} \pi_i}}{\frac{1-\sum_{i=1}^{K-1} \pi_i + \sum_{i=1}^{K-1} \pi_i}{1-\sum_{i=1}^{K-1} \pi_i}} = M \frac{\frac{\pi_i}{1-\sum_{i=1}^{K-1} \pi_i}}{\frac{1}{1-\sum_{i=1}^{K-1} \pi_i}} = M\pi_i$$

$$\begin{aligned} \mathrm{Cov}(x_i, x_j) = \frac{\partial^2 \mathcal{A}(\eta)}{\partial \eta_i \partial \eta_j} &= -M \frac{\exp(\eta_i + \eta_j)}{(1 + \sum_{i=1}^{K-1} \exp(\eta_i))^2} \\ &= -M \frac{\frac{\pi_i}{1-\sum_{i=1}^{K-1} \pi_i} \cdot \frac{\pi_j}{1-\sum_{i=1}^{K-1} \pi_i}}{\frac{1}{1-\sum_{i=1}^{K-1} \pi_i} \cdot \frac{1}{1-\sum_{i=1}^{K-1} \pi_i}} \\ &= -M\pi_i\pi_j \end{aligned}$$

Problem 3

1.

The given model is a linear noisy ICA model: We assume underlying source distributions that are **not Gaussian** (i.e. our $s_{it} \sim \mathcal{T}(0, \nu_i)$ are assumed to be student-t distributed) that are **linearly combined into our observed signals** x with

$$x_{kt} = \sum_{k=1}^{K_S} A_{ki} s_{it} + \epsilon_{kt} \qquad \epsilon \sim \mathcal{N}(0, \sigma_k^2)$$

where A is the mixing matrix that combines the **independent** underlying source signals s linearly and ϵ is some noise.

2.

$$\begin{aligned} p(\{s_1\}, \{s_2\}, \{x_1\}, \{x_2\}, \{x_3\}) &\overset{\text{i.i.d.}}{=} \prod_t^T p(\{s_{1t}\}, \{s_{2t}\}, \{x_{1t}\}, \{x_{2t}\}, \{x_{3t}\}) \\ &= \prod_t^T p(s_{1t}) p(s_{2t}) p(x_{1t}|s_{1t}, s_{2t}) p(x_{2t}|s_{1t}, s_{2t}) p(x_{3t}|s_{1t}, s_{2t}) \\ &= \prod_t^T \ \mathcal{T}(s_{1t}|0, \nu_1) \mathcal{T}(s_{2t}|0, \nu_2) \mathcal{N}(x_{1t}|A_1, \sigma_1, s_{1t}, s_{2t}) \\ &\qquad \mathcal{N}(x_{2t}|A_2, \sigma_2, s_{1t}, s_{2t}) \mathcal{N}(x_{3t}|A_3, \sigma_3, s_{1t}, s_{2t}) \end{aligned}$$

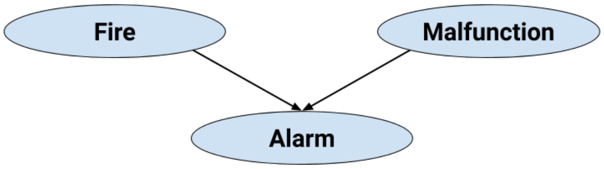

Figure 1: Bayesian network with 3 variables. ”Explaining away” of either **Fire** or **Malfunction** can happen if the value of **Alarm** and the respective other variable is given.

3.

Consider the Bayesian network given in figure 1: The two random variables **Fire** and **Malfunction** are independent from each other in the case that the value of the random variable **Alarm** is not given. However, if the value of **Alarm** is given, then the two originally independent variables **Fire** and **Malfunction** become dependent. In the case that we then also know about the value of either **Fire** or **Malfunction**, this given value influences the likeliness of the value for the respective other random variable. One specific case of this phenomenon is referred to as ”explaining away”, where the observation for one of the initially independent variables makes the other less likely. In the given graph, an example would be if there would be an alarm and we would know that there is an malfunction. In this case, even though we observe an alarm, we would think of a fire as less likely than if we would if we didn’t know about the existing malfunction.
The same thing can happen in the given ICA model: Although the source signals s_{1t} and s_{2t} are independent from each other, they can ”explain away” each other in the case that at least one $x_{it} \quad i \in \{1, 2, 3\}$ is given.

4.

(a) False

(b) True

(c) False

(d) True

(e) False

(f) False

(g) False

(h) False

5.

Since we know that Markov blankets include parents, co-parents and children:

Markov blanket of s_1 : $\{s_2, x_1, x_2, x_3\}$

Markov blanket of x_1 : $\{s_1, s_2\}$

6.

$$\begin{aligned} p(\{x_{kt}\}|W, \{\nu_i\}) &= \prod_{t=1}^{T} p_X(x_t) = \prod_{t=1}^{T} p_S(s(x_t)) \cdot |\det \frac{\partial s}{\partial x}| \\ &= \prod_{t=1}^{T} p_S(Wx_t) \cdot |\det W| = \prod_{t=1}^{T} |\det W| \prod_{i=1}^{K} \mathcal{T}(W_i x_t|0, \nu_i) \end{aligned}$$

7.

$$\log p(\{x_{kt}\}|W, \{\nu_i\}) = \sum_{t=1}^{T} \Big(\log(|\det W|) + \sum_{i=1}^{K} \log(\mathcal{T}(W_i x_t|0, \nu_i)) \Big)$$

9.

Since the model parameters are represented by the "unmixing matrix" $W \in \mathbb{R}^{K \times K}$ I would expect the model to overfit in the limit $K >> T$. In this case we have much more parameters than observed time points for each recording, which means the observed data can be overly good represented with the model (== overfitting).